HADOOP COURSE CONTENT – (HADOOP-1.X, 2.X & 3.X)

(Development, Administration & REAL TIME Projects Implementation)

Introduction to BIGDATA and HADOOP

- What is Big Data?
- What is Hadoop?
- Relation between Big Data and Hadoop.
- What is the need of going ahead with Hadoop?
- Scenarios to apt Hadoop Technology in REAL TIME Projects
- Challenges with Big Data
    - ➢ Storage
    - ➢ Processing
- How Hadoop is addressing Big Data Changes
- Comparison with Other Technologies
    - ➢ RDBMS
    - ➢ Data Warehouse
    - ➢ TeraData
- Different Components of Hadoop Echo System
    - ➢ Storage Components
    - ➢ Processing Components
- Importance of Hadoop Echo System Components in Real Time Projects
- Other solutions of Big Data
    - ➢ Introduction to NO SQL
    - ➢ NO SQL vs HADOOP
- Type of BigData Projects
    - ➢ On Premises project
    - ➢ Cloud Integrated Project
    - ➢ Differences between On Premises & Cloud Integrated Projects

HDFS (Hadoop Distributed File System)

- What is a Cluster Environment?
- Cluster Vs Hadoop Cluster.
- Significance of HDFS in Hadoop
- Features of HDFS
- Storage aspects of HDFS
    - ➢ Block – the basic storage unit in hadoop
    - ➢ How to Configure block size
    - ➢ Default Vs Configurable Block size
    - ➢ Why HDFS Block size so large?
    - ➢ Design Principles of Block Size

- HDFS Architecture - 5 Daemons of Hadoop
    - NameNode and its functionality
    - DataNode and its functionality
    - JobTracker and its functionality
    - TaskTrack and its functionality
    - Secondary Name Node and its functionality.

- Replication in Hadoop – Fail Over Mechanism
    - Data Storage in Data Nodes
    - Fail Over Mechanism in Hadoop – Replication
    - Replication Configuration
    - Custom Replication
    - Design Constraints with Replication Factor
    - Can we change the replication factor in Hadoop?
    - Can we change the block size for a file or directory in Hadoop?

- Accessing HDFS
    - CLI (Command Line Interface) and HDFS Commands
    - Java Based Approach
- Hadoop Archives
- Configuration files in Hadoop Installation and the Purpose
- How to & Where to Configure Hadoop Daemons in a Hadoop Cluster?
- Difference between Hadoop 1.X.X , Hadoop 2.X.X & 3.X.X version
    - Name Node HA (High Availability in Hadoop 2.X.X)
    - Importance of NFS in Hadoop-2.X
    - Importance of Journal Nodes in Hadoop-2.X

MapReduce

- **Why Map Reduce is essential in Hadoop?**
- **Processing Daemons of Hadoop**
    - ➢ **Job Tracker**
        - ✓ Roles Of Job Tracker
        - ✓ Drawbacks w.r.to Job Tracker failure in Hadoop Cluster
        - ✓ How to configure Job Tracker in Hadoop Cluster
    - ➢ **Task Tracker**
        - ✓ Roles of Task Tracker
        - ✓ Drawbacks w.r.to Task Tracker Failure in Hadoop Cluster

- **Input Split**
    - ✓ InputSplit
    - ✓ Need Of Input Split in Map Reduce
    - ✓ InputSplit Size
    - ✓ InputSplit Size Vs Block Size

- ✓ InputSplit Vs Mappers

- **Map Reduce Life Cycle**
  - ✓ Communication Mechanism of Job Tracker & Task Tracker
  - ✓ Input Format Class
  - ✓ Record Reader Class
  - ✓ Success Case Scenarios
  - ✓ Failure Case Scenarios
  - ✓ Retry Mechanism in Map Reduce

- **MapReduce Programming Model**
  - **Different phases of Map Reduce Algorithm**
  - **Different Data types in Map Reduce**
    - ✓ Primitive Data types Vs Map Reduce Data types
  - **How to write a basic Map Reduce Program**
    - Driver Code
    - Mapper Code
    - Reducer Code

  - **Driver Code**
    - ✓ Importance of Driver Code in a Map Reduce program
    - ✓ How to Identify the Driver Code in Map Reduce program
    - ✓ Different sections of Driver code

  - **Mapper Code**
    - ✓ Importance of Mapper Phase in Map Reduce
    - ✓ How to Write a Mapper Class?
    - ✓ Methods in Mapper Class

  - **Reducer Code**
    - ✓ Importance of Reduce phase in Map Reduce
    - ✓ How to Write Reducer Class?
    - ✓ Methods in Reducer Class
  - **IDENTITY MAPPER & IDENTITY REDUCER**
  - **Input Format's in Map Reduce**
    - ✓ TextInputFormat
    - ✓ KeyValueTextInputFormat
    - ✓ NLineInputFormat
    - ✓ DBInputFormat
    - ✓ SequenceFileInputFormat.
    - ✓ How to use the specific input format in Map Reduce
    - ✓ How to write Custom Input Format Class and Custom Record Reader
  - **Output Format's in Map Reduce**

- ✓ TextOutputFormat
- ✓ KeyValueTextOutputFormat
- ✓ NLineOutputFormat
- ✓ DBOutputFormat
- ✓ SequenceFileOutputFormat.
- ✓ How to use the specific Output format in Map Reduce
- ✓ How to write Custom Output Format Class and Custom Record Writer

- **Map Reduce API(Application Programming Interface)**
  - ✓ New API
  - ✓ Deprecated API

- **Combiner in Map Reduce**
  - ✓ Is combiner mandate in Map Reduce
  - ✓ How to use the combiner class in Map Reduce
  - ✓ Performance tradeoffs w.r.to Combiner
  - ✓ Real Time Use Cases
  - ✓ Where to Use & Where Not to Use Combiner

- **Partitioner in Map Reduce**
  - ✓ Importance of Practitioner class in Map Reduce
  - ✓ How to use the Partitioner class in Map Reduce
  - ✓ Different types of Practitioners in Map Reducer
  - ✓ Importance of hashPartitioner
  - ✓ How to write a custom Practitioner
  - ✓ Real Time Use Cases

- **Compression Techniques in Map Reduce**
  - ✓ Importance of Compression in Map Reduce
  - ✓ What is CODEC
  - ✓ Compression Types
  - ✓ GzipCodec
  - ✓ BzipCodec
  - ✓ LZOCodec
  - ✓ SnappuCodec
  - ✓ Configurations w.r.to Compression Techinques
  - ✓ How to customize the Compression per one job Vs all the job.

- **Map Reduce Job Chaining**
  - ✓ What is Map Reduce Job Chaining?
  - ✓ Use of MR Chaining in Real Time Hadoop Projects
  - ✓ Real Time Use case
  - ✓ Performance trade off's using MR Chaining

- - **Joins - in Map Reduce**
    - ✓ Map Side Join
    - ✓ Reduce Side Join
    - ✓ Performance Trade Off
    - ✓ Real Time applicability of Map Side & Reduce Side Joins in Map Reduce
    - ✓ Distributed cache
  - **How to debug MapReduce Jobs in Local and Pseudo cluster Mode.**
  - **Introduction to MapReduce Streaming**
  - **Data locality in Map Reduce**
- **Secondary Sorting Using Map Reduce**

Apache PIG

- Introduction to Apache Pig
- Map Reduce Vs Apache Pig
- SQL Vs Apache Pig
- Different datat ypes in Pig
- Where to Use Map Reduce and PIG in REAL Time Hadoop Projects
- Modes Of Execution in Pig
    - ✓ Local Mode
    - ✓ Map Reduce OR Distributed Mode
- Execution Mechanism
    - ✓ Grunt Shell
    - ✓ Script
    - ✓ Embedded
- Transformations in Pig
- How to write a simple pig script
- Parameter substitution in PIG Scripts
- XML Processing through PIG
- JSON Processing through PIG
- Importance of DEFINE Keyword in PIG
- How to develop the Complex Pig Script
- Bags , Tuples and fields in PIG
- UDFs in Pig
    - ✓ Need of using UDFs in PIG
    - ✓ How to use UDFs
    - ✓ REGISTER Key word in PIG
- Techniques to improve the performance and efficiency of Pig Latin Programs

HIVE

- Hive Introduction

- Need of Apache HIVE in Hadoop
- When to choose MAP REDUCE , PIG & HIVE in REAL Time Project
- Hive Architecture
    - Driver
    - Compiler
    - Executor(Semantic Analyzer)

- Meta Store in Hive
    - Importance Of Hive Meta Store
    - Embedded Metastore VS External Metastore
    - Embedded metastore configuration
    - External metastore configuration
    - Communication mechanism with Metastore and configuration details
    - Drawbacks with Internal/Embedded metastore over External metastore

- Hive Integration with Hadoop
- Hive Query Language(Hive QL)
- Configuring Hive with MySQL MetaStore
- SQL VS Hive QL
- Data Slicing Mechanisms
    - Partitions In Hive
        - Static Partitioning in Hive and its performance trade offs
        - Dynamic Partitioning in Hive and its performance trade offs
    - Buckets In Hive
    - Partitioning with Bucketing usage in Real Time Project Use Cases
    - Partitioning Vs Bucketing
    - Real Time Use Cases

- Collection Data Types in HIVE
    - Array
    - Struct
    - Map
    - Real Time Use Cases
- Conditional Functions in HIVE
    - Imporatnce of CASE Statement
    - Real Time Use Cases on CASE Statements
    -
- DATE Functions in HIVE
    - Imporatnce of Date Functions
    - Real Time Use Cases on DATE Functions

- User Defined Functions(UDFs) in HIVE
  - ✓ UDFs
  - ✓ UDAFs
  - ✓ UDTFs
  - ✓ Need of UDFs in HIVE
- Hive Serializer/Deserializer - SerDe
- Semi Structured Data Processing Using Hive
- Semi Structured Data Processing through HIVE
  - ✓ XML Data Processing
  - ✓ Importance of XML Data Processing through HIVE in Real Time Projects
  - ✓ JSON (Java Script Object Notation) Data Processing through HIVE
  - ✓ Importance of JSON Data Processing through HIVE in Real Time Projects
  - ✓
- HIVE – HBASE Integration
  - ✓ Importance of HIVE – HBASE Integration with respect to Latency
  - ✓ Real Time Use Cases on Hive – HBase Integration

SQOOP

- Introduction to Sqoop.
- MySQL client and Server Installation
- How to connect to Relational Database using Sqoop
- Performance Implications in SQOOP Import and how to improve the performance
- Performance Implications in SQOOP Export and how to improve the performance
- Different Sqoop Commands
  - Different flavors of Imports
  - Export
  - Hive-Imports
- SQOOP Incremental Load VS History Load & Limitations in Incremental Load

HBase

- Different BigData Solutions - Hadoop Comparision with Not Only SQL(NO SQL)
- Hbase introduction

- HDFS Vs HBase
- HBase Vs RDBMS
- HBase Vs Cassandra VS Mongo DB & Real Time Use Cases on applicabiltiy
- Hbase usecases
- Hbase Data modeling Elements
  - Column families
  - Column Qualifier Name
  - Row Key
- HBase Architecture
- Bulk Loading Operation with HBASE
  - Importance of **ImportTsv** Utility in HBase
  - Real Time case study on the usage of **ImportTSV** Utility of HBase
- Clients
  - REST
  - Thrift
  - Java Based
  - Avro
- Map Reduce Integration
- Map Reduce over HBase
- HBase Admin
  - Schema Definition
  - Basic CRUD Operations
  - Client Side Buffering in HBase

Flume

- Flume Introduction
- Flume Architecture
- Flume Master , Flume Collector and Flume Agent
- Flume Configurations
- Real Time Use Case using Apache Flume
- Sentimental Data Analytics with respect to Social Media Data with Flume & Hive

Oozie

- Oozie Introduction
- Oozie Architectrure
- Oozie Configuration Files
- Oozie Job Submission
  - ✓ Workflow.xml
  - ✓ Coordinator.xml

- ✓ job.coordinator.properties
- ✓ Transit parameters in workflow.xml

YARN (Yet another Resource Negotiator) – Next Gen. Map Reduce

- What is YARN?
- Difference between Map Reduce & YARN
- YARN Architecture
  - ✓ Resource Manager
  - ✓ Application Master
  - ✓ Node Manager
- When should we go ahead with YARN
- YARN Process flow
- YARN Web UI
- Different Configuration Files for YARN
- How to access Map Reduce Job History Server and Importance of Historyserver
- Examples on YARN

Cloudera Impala

- What is Impala?
- How can we use Impala for Query Processing?
- When should we go ahead with Impala
- Data Analytics with respect to Hive Batch Processing VS Impala Real Time Processing
- REAL TIME Use Cases with Impala

MongoDB (As part of NoSQL Databases)

- Need of NoSQL Databases
- Relational VS Non-Relational Databases
- Introduction to MongoDB
- Features of MongoDB
- Installation of MongoDB
- Mongo DB Basic operations
- REAL Time Use Cases on Hadoop Data Processing & MongoDB Storage

Apache Cassandra

- Introduction to Cassandra
- Mongo DB Vs Cassandra
- Basic Operation using Cassandra
- Comparison among HBase , Mongo DB and Cassandra NO SQL DBs

Apache Kafka (A Distributed Message Queuing System)

- Introduction to Kafka
- Installation of Kafka
- Difference between MQ Vs Kafka
- Basic Operation using Kafka and real time case study on Kafka usage

Mahout (As a part of BIGDATA ANALYTICS)

- Introduction to Machine Learning (ML) Languages
- Types of Machine Learning
- Introduction to Apache MAHOUT
- Categories of Mahout Algorithms
- Real Time Use case using Classifier Algorithm of Mahout – Naives Bayes

Apache Spark – with Scala Content
[As part of Hadoop Course]

Introduction to SCALA

- Why Scala
- Scala Vs Java
- Why Scala is a Hybrid Language
- Pre-Requisits for Scala Installation

SCALA Basics

- Scala Data types
- Scala Packages
- Runtime environment of Scala & Java
- Different IDE Support for Scala
- Control Structures

Interactive SCALA – SCALA Shell

- Scala REPL [Real Evaluate Print Loop]
- Writing Scala Scripts
- Compiling the Scala Programs
- Different IDEs for Scala

SCALA Type Less, Do More

- Var[variable] VS val[Value]
- Type Inference
- DataTyes in SCALA
- Type Casting in Scala

Conditional Statements in SCALA

- If expression
- If-else expression
- While Loop and Do…While Loop & difference between the two
- For loop , different forms of for loop in SCALA
- Pattern matching in SCALA & use of **case** and **match** keywords in SCALA

Functional Programing in SCALA

- What is Functional Programming
- Difference between Object Oriented and Functional Programing Paradigm
- Closures in Scala
- Currying Functions in Scala
- Higher Ordered Functions in Scala

SCALA Environment Set Up

- Scala set up on Linux
- Java Set Up
- Scala Set Up

SCALA Collections

- List
- Set
- Map

SCALA Object Oriented Programming Introducton

SPARK

- **Introduction to Spark**

  - Motivation for Spark
  - Spark Vs Map Reduce Processing
  - Architecture Of Spark
  - Spark Shell Introduction
  - Creating Spark Context
  - File Operations in Spark Shell
  - Caching in Spark
  - Real time Examples of Spark
  - Introduction to Spark Components
    - ✓ Spark Core
    - ✓ Spark SQL

- ✓Spark Streaming
- ✓Spark MLLib
- ✓Spark Streaming

- **Spark Core**

Resilient Distributed Dataset [RDD]

- What is RDD and why it is important in Spark
- Core Features of RDD
  1. Lazily Evaluated
  2. Immutable
  3. Partitioned
- Different Operation on RDDs
  1. Transformations
  2. Actions
- Transformation in RDD
- Different Examples on Transformations
- Actions in RDD
- Different examples on Actions
- Loading Data through RDD
- Saving Data
- Key-Value pair RDD
- Pair RDD operations
- Running Spark in a Clustered Mode
- Deploying Application with spark-submit
- Cluster Management

Spark SQL

- Introduction to Spark SQL
- The SQL Context
- Hive Vs Spark SQL
- Introduction to Data Frames [DFs]
- Examples on Spark SQL

Different File Formats Processing through Spark SQL

- CSV
- JSON
- PARQUET
- ORC
- TEXT

Spark SQL Integrations

- Spark – Hive Integration and Real Time use cases on the same
- Spark – RDBMS Integration and Real Time use cases on the same
- Spark – NO SQL Integration Introduction and Importance

Introduction to Big Data Project Integration with AWS Cloud

HADOOP ADMINISTRATION TOPICS

- Hadoop Single Node Cluster Set Up (Hands on Installation on Laptops)
    - ✓ Operating System Installation
    - ✓ JDK Installation
    - ✓ SSH Configuration.
    - ✓ Dedicated Group & User Creation
    - ✓ Hadoop Installation
    - ✓ Different Configuration Files Setting
    - ✓ Name node format
    - ✓ Starting the Hadoop Daemons
- Multi Node Hadoop Cluster Set Up (Hands on Installation on Laptops)
    - ✓ Network related settings
    - ✓ Hosts Configuration
    - ✓ Password less SSH Communication
    - ✓ Hadoop Installation
    - ✓ Configuration Files Setting
    - ✓ Name Node Format
    - ✓ Starting the Hadoop Daemons
- PIG Installation (Hands on Installation on Laptops)
    - ✓ Local Mode
    - ✓ Clustered Mode
    - ✓ Bashrc file configuration
- SQOOP Installation (Hands on Installation on Laptops)
    - ✓ Sqoop installation with MySQL Client
- HIVE Installation(Hands on Installation on Laptops)
    - ✓ Local Mode
    - ✓ Clustered Mode
- HBase Installation (Hands on Installation on Laptops)
    - Local Mode
    - Clustered Mode

- OOZIE Installation (Hands on Installation on Laptops)
- Mongo DB Installation (Hands on Installation on Laptops)
- SPARK Installation (Hands on Installation on Laptops)
- SCALA Installation (Hands on Installation on Laptops)

- Commissioning Of Nodes In Hadoop Cluster
- Decommissioning Of Nodes from Hadoop Cluster

PRE-REQUISITES FOR THE COURSE

- ➢ SQL Commands Basic Knowledge [FREE SQL Classes will be provided as part of the course itself]
- ➢ Linux Basic Commands [FREE Classes provided as part of course]
- ➢ Java Basics - OOPs Concepts only [FREE Java OOPs Concept Classes will provided as part of course]

What we are offering as part of the Course?

- 3 REAL TIME Hadoop Projects End-to-End Explanation with architecture.
- End to End Hadoop Real Time Project derivation workshop.
- FREE Online Mock Test based on CCA 175 exam.
- Mock Interviews will be conducted on a one-to-one basis after the course duration.
- Hand Written Hard Copy & Soft Copy Materials for all the Components.
- Detailed Assistance in RESUME Preparation on a one-to-one basis with Real Time Projects based on your technical back ground.
- All the Real time interview questions and answers will be provided.
- 15 classes for Core Java, Linux and SQL concepts will be covered as part of the course.
- Discussing the new happenings in Hadoop
- FREE BIGDATA workshops
- Discussing the Interview Questions on a daily basis
- Discussing Certification (CCA 175 – Spark and Hadoop Certification) Related topics on a daily basis.
- 2 Written Exams will be conducted during the course with Real Time Scenarios
  Which will help a lot where you stand with market standands?
- Academic Projects will be provided for pursuing students.